CODIAC-D-12
11 July 1958

I N T E L L I G E N C E   A D V I S O R Y   C O M M I T T E E

MEMORANDUM FOR: IAC Committee on Documentation

SUBJECT : Translation of Russian Article on Automation.

The attached article is forwarded for your information pursuant to the discussion at the CODIAC meeting of 9 July 1958. It is the earliest systematic treatment of the mechanization of scientific information thus far identified in the Russian literature. Subsequent events indicate that Dr. Gutenmakher is a leading if not the key figure in Russian research in this field.

[redacted] 25X1
Secretary

Attachment
As stated above.

Approved For Release 2004/03/31 : CIA-RDP80B01139A000200010003-8

CODIAC-D-12
11 July 1958

INTELLIGENCE ADVISORY COMMITTEE

COMMITTEE ON DOCUMENTATION

The Problem of the Mechanization of Scientific Information

by

I. I. Gutenmakher, Doctor of Technical Sciences

From: Vestnik Akademii Nauk, SSSR, No. 8, (Aug. 1952) pp. 46-52

The problem of the mechanization of scientific information, posed by Academician A. N. Nesmeyanov, President of the Academy of Sciences of USSR, is a knotty one, the solution of which must bring about a fundamental chan[ge] in the methods of handling scientific-technical information, will contribu[te] to the optimum organization of scientific work, and will assist in the mos[t] rapid introduction of the achievements of science and technology into the national economy.

The continual increase in the number of scientific investigations and technical developments has led to a torrential flow of literature, which ha[s] literally engulfed scientists and engineers.

The total number of printed works accumulated by humanity is very great. It is estimated to be of the order of hundreds of millions. At the present rates of increase of the quantity of literature, the contents of libraries will be almost doubled every 10 - 15 years. In 50 - 60 years an increase of the contents of libraries to 15 - 20 times of their present size may be expected. Specialists in any particular field of knowledge cannot follow the progress in allied fields of science and technology. For systematization and selection of literature, a large army of bibliographers compiles surveys which aid specialists.

The present practice requires the solution of complex technical and scientific problems in minimal time with recognition of all accumulated data. Moreover, a substantial part of the time of scientists and specialists is spent in the selection of literature and in obtaining exhaustive information on their problems.

Because of the abundance of data scattered in a vast number of journals, books, and reports, a search for the necessary information sometimes requires so much time and work that it is easier to repeat either the experiment or the calculation than to find its description in the literature.

Attempts to classify informational material by means of various systems of library classification cannot lead to an efficient solution of the question.

Classification, as a method of organized disposition of materials, is characterized by selection of certain criteria as a basis for division of information into independent "non-overlapping" groups. Thus, for example, bacteria are classified according to their morphology, pathological activity, conditions governing their lives and growth, etc.

Chemical compounds are classified on the basis of their composition, chemical structure, physical properties, and according to their practical application. Electronic apparatus is classified according to the field of application, power, performance, etc.

Classification suggests principally the variation of certain criteria. It is impossible to utilize all possible combinations of all the criteria.

The attempt to create such a detailed and finely divided classification was destined to failure so the classification had to be accomplished by the simplest possible questions.

The theory of combination easily permits the determination of those "astronomical" figures, which are obtained by calculation of the possible number of elementary questions in such a system. This number, even for ordinary raw data, will be of the order of 10 to the thousandth power (10^{1000}).

Connected with these problems is the review and selection of literature by every individual scientist and specialist. Help in this work is provided by bibliographers, special departments in libraries, institutes, ministries, and services.

For example, calculation has been attempted of the amount of work necessary for the satisfaction of daily requirements for scientific information.

In the Soviet Union there are several million engineers, technicians, scientific workers, etc. Let each of them require scientific information only once a year (selection of literature concerning an interesting question). Let us consider that in one year this constitutes about 3 - 6 million inquiries, or 10 - 20 thousand inquiries per day.

Let us assume that the satisfaction of each requirement for information is derived from material, the volume of which on the average is limited to one thousand pages of text. Let us note also that one person scans 100 pages of text in a day. In this case then, for each selection 10 man-days are required.

Thus, for the completion of all of the work the labor of 100 - 200 thousand qualified readers is necessary, looking through material in accord with a large number of requirements.

We shall work from the desired scale of informational work, and not from the present situation. But even if these figures are reduced, even then they will indicate the large volume of scientific-information work necessary for a country.

The continually rising scientific level of Soviet specialists requires a procedure for scientific-information work which would provide each of them on the average not one, as is assumed in our calculation, but several pieces of information per year.

It may be considered that because of insufficient information a significant part of the effort of the staff and facilities of scientific institutions is wasted on the repetition of investigations which have already been carried out. Much time is used in the selection of information at the beginning of any large-scale investigation, because before beginning a new scientific research, it is necessary to become familiar with data from the total literature. Paraphrasing V. V. Mayakovskiy, it can be said that a scientist, gathering scientific-information material, investigates a thousand tons of "word-ore" for a single nugget of information.

Information material accumulated in libraries is a vast potential wealth which brings all the more benefit the better organized is the scientific-information work.

J. V. Stalin indicates that the mechanization of work process is that determining strength, without which it is impossible to maintain either our tempo or new scale of production. This point can fully apply to the arrangement of scientific information.

Mechanization by no means is limited just to the simple acceleration of the selection of material. Insofar as research may be directed not only toward solution of Question A, occurring together with Question B, but also in the solution of B according to A, then this relation can help in the establishment of connections of the type of bond between cause and effect.

The possibility of a quick abstract of accumulated data in accord with specific questions leads, moreover, to weakening, and even to the elimination of the rapidly growing divisions of science. At present, specialists, even of allied disciplines, understand one another with difficulty. Major difficulties are eliminated by searching for and utilization of analogies in phenomena, processes and structures found in different fields. The preparation of material for informational-bibliographical machines uses generalizations of results of the most diverse investigations and developments.

According to an idea of A. N. Nesmeyanov, it is necessary to be able mechanically to look through the contents of information in accordance with given requirements, thus resulting in the selection of required material from a number of independent references.

Thus, the contents of every work must be expressed in an abstract of a certain number of the simplest, elementary propositions ---- theses, facts, propositions and criteria. Scientific hypotheses, conceptions, results of experiments, constructions, principles of the functioning of apparatus, physical constants, time, place of action and other informational data must be concisely and simply formulated.

In first approximation, we can assume that an abstract of an average journal article will contain 100 - 200 such propositions.

The analysis of all incoming material, the formulation and statement of elementary propositions can be produced, according to appropriate rules and specifications, by the authors of the articles or by special workers. A large part of this work can be accomplished by the newly organized Institute of Scientific Information by preparation of abstracts according to the different divisions of science and engineering.

The selection of accumulated material for a given question must be done mechanically.

Questions must also be put in the form of the simplest elementary propositions. A large number of such questions can be made simultaneously. During the search of the information, the machine must select only those pieces of information which simultaneously are pertinent to all the questions posed.

The answer must contain a list of selected works (in the form of a list of issues of these works in a bibliographical order) and give their contents.

The problem of obtaining photostats or originals of the selected works must be solved separately. This problem may be solved also by means of automatization (photoautomats and other apparatus).

The fundamental problem of an automatic device (for brevity we shall designate this device —— machine) —— is the selection of a bibliography according to a combination of a number of requirements (questions). With a large number of these requirements, the number of possible combinations is practically infinite.

In certain cases, of course, there can be a combination, which is not reflected in any of the materials. A negative answer to such a question is also useful because it testifies to the newness of the problem to research or development.

Insofar as the selection is controlled by the contents of material, the machine, according to the design, can answer the most varied questions in any combination of given requirements.

Let us select and process, for example, information about the physical constants of molecules. The problem for the machine obviously will be formulated in the following way: to find, by annual volume of issues, publications in which certain constants of molecules of determined composition are important and within certain limits of variation. The number of criteria may vary from one to an infinite number.

However, the possibilities of a machine, built according to these principles are very broad. In a number of the inquiries, first interest may not be in a bibliography, but rather in an analysis of the very contents of the publications.

Let us assume, for example, the development of informational material concerning chemical kinetics and data on the mechanism of chemical reactions.

The tasks of mechanical selection of information can be formulated as follows:

1. to find papers in which slow reactions are discussed, i.e., those reactions where the pre-exponential term is important within defined limits.
2. to find papers, in which is indicated that a certain reaction proceeds with the velocity specified in the question and also data concerning temperature and concentration.

Thus, it is possible to set requirements, the numerical value of certain criteria connected with them, and to require a report of other associated information (of all or part of it).

At first it may seem that the limitation of an answer by just a certain set of criteria without an indication of source is meaningless. But if it is remembered what a huge quantity of material may be examined mechanically, then the expediency and effectiveness of such a method of utilization of the machine become evident.

A matter often cannot wait and may require an urgent answer to a question concerning relationships of certain quantities, correspondence or discrepancy of a series of criteria, etc. Thus, for example, in the case indicated above, it is interesting to check whether there are inconsistent data for corresponding reactions under the same conditions or under different conditions (for example a reaction in a temperature range from t_1 to t_2 conforms to a usual kinetic equation, but from t_2 to t_3 the presence of a more complicated chain reaction is indicated).

The system of mechanization permits one quickly and easily to extract accumulated knowledge to compare different factors, to analyze data, etc.

However, for the realization of this system it is necessary to solve many difficult problems.

1. <u>The Creation of an Efficient Well-Defined System of Recording of Information.</u>

Science, having generated this problem, has also prepared the means for its solution. It is sufficient, for example, to remember the existence of chemical formulae, which characterize the structure of substances. The theory of dimensions permits one to express different physical values by means of a small number of basic quantities. Thus, for the characteristics of phenomena studied in mechanics, length (L), mass (M) and time (T) may be assumed as bases, and then the dimension of mechanical values appears in the following form: force --- (L^2MT^{-2}), velocity --- (LT^{-2}), density --- (ML^{-3}), force --- (L^2MT^{-3}), etc.

For the characteristics of electromagnetic phenomena to these basic values is added the fourth value --- dielectric strength (ε) or magnetic permeability (μ).

Thus the words "electrical field intensity" may be written by the formula of dimension ($L^{-\frac{1}{2}}M^{\frac{1}{2}}T^{-1}\varepsilon^{\frac{1}{2}}$). If in a text, information is expressed in terms of "electromotive force", "potential", "tension", "voltage" (the latter term is eliminated), then they may all be expressed identically by the formula: ($L^{\frac{1}{2}}M^{\frac{1}{2}}T^{-1}\varepsilon^{-\frac{1}{2}}$).

The formation of a unique dictionary which generalizes the now existing numerous specialized dictionaries would permit one to find information in the most unexpected places.

There are numerous well-known examples, when "new discoveries" in one field have long been utilized in another. For example, the negative feedback in the mechanical amplifier of a governor of a steam engine has been used to increase stability for about 80 years, but for electronic tube amplifiers, it was "rediscovered" just in the 1930's, and only 5 years later --- for magnetic amplifiers.

The tendency toward maximum generalization exists in every field of science. Significant achievements in this direction were made, for example, in the theory of oscillation developed by Soviet physicists. For the generalization of material the experience of the theory of similarity and analogy of physical phenomena may be applied with great advantage.

The method of analogies is based on the condition which V. I. Lenin brilliantly observed --- "The unity of nature appears in "striking analogy" of differential equations pertaining to different fields of phenomena".

The mathematical analogies of mechanical, acoustic, hydraulic and other phenomena are widely known. A. N. Krylov indicated that such "analogies between problems of totally different fields, but leading to identical differential equations, may be numerous. Would it have seemed possible to have found anything in common between a calculation of movement of heavenly bodies under the effect of attraction to the sun and among themselves and the rolling of a ship on the heaving sea, or between the computation of the so-called secular inequalities in the movement of heavenly bodies and the torsional oscillations of the drive shaft of a multicylinder diesel engine operating on the ship propeller or on an electric generator? Meanwhile, if such a formula and equations without words can be written then it is impossible to distinguish which of these questions is being resolved: the equations are one and the same".

Consequently, the presence of analogies in these diverse physical phenomena permits one to describe them in the following form:

a. by a system of general equations

b. by dimensional formulae of the values found

c. by dimensionless values (by criteria of similarity)

d. by a series of elementary propositions, indicating the purpose and the results of the research or developments.

In many cases the most precise and concise formulations may be obtained by means of mathematical formulae. This symbolic economical form of recording different concepts has undergone a great change and continues to develop.

For example, the recording of an algebraic equation $x^3 + ax = b$ 400 years ago would have appeard thus X cubus + A planum X acquator B solido.

At present there is an apparatus for symbolic recording of the most complicated logical conceptions, operations and conclusions (for example, algebra of logic).

The utilization of the arsenal of mathematical means gives excellent results. The experience from the theory of similarity, the introduction of dimensionless quantities (criteria of similarity) for evaluation of the quantities found, in comparison with single basic units is also very useful.

We should recall the great experience gained in the clear formulation of statements in patent matters. It is well known that it is necessary to write the formula of an invention in the form of a list of a number of different elementary propositions, in which each proposition is expressed, so to speak, "without drawing a breath".

If all these efforts in different fields were added up and developed, then the result obtained will have great independent scientific significance.

The development of the technique of scientific information, according to the ideas of A. N. Nesmeyanov, requires the creation of a theoretical basis for the generalization of information which logically is the next stage of the development of a theory of similarity and analogy of phenomena. The successful solution of this problem will lead to a still more efficient utilization of the position of dialectical materialism concerning the reflection of the unity of nature in the development of science, to the strengthening of the common and interconnected bonds among the different fields of science.

2. <u>The Development of a Rational System of Classification of Material.</u>

If the recording system answers the question of what to look for, then the system of classification indicates where to look for the material.

If the machine method permitted, for example, a time of the order of 10 - 20 minutes "to look through" all of the accumulated material, then, in general, the task of the creation of a system of classification would recede.

Regrettably, the quantity of informational data is so great, that with the greatest speed of review it will be impossible to look through all of the material in 10 - 20 minutes.

Two preliminary calculations show that if during one year 100,000 abstracts were entered and each of them were on the average, formulated of 100 elementary sentences of 10 words each, then this would amount to 100 million words a year. For the review of such a quantity of material in 10 minutes is required around 1/100,000 of a second per word.

For the review of material of 10 years input, the speed of scan would have to be increased still more and brought up to a millionth part of a second per word, or the time of scan would have to be increased.

It is understood, in many of the important cases, to avoid undesirable omission of material, it perhaps will be necessary to increase time and require scanning of all the accumulated information according to given criteria.

In the majority of cases the scope of scan may be limited. If information concerning cutting tools of lathes is required, then it is hardly worth looking for it in literature concerning electronic oscillographs. Therefore, it will be expedient to introduce some flexible system of classification for facilitating and accelerating searches of material.

In this connection there would probably have to be worked out a new variation of classification for a period of time, during which it will be expedient to find some general method of indexing material.

Classification by machine --- this system of addresses in a commutator must be constructed so that it would be relatively easy to change with a change of classification.

Inasmuch as the machine method is a fast method of search, the system must be relatively "consolidated". The structural plan of classification will, probably, have the form of "a tree of knowledge" with a different number of "branches" in the divisions.

In programming a search of information, it is possible to list a relatively large number of "addresses" of those "branches". In which the presence of material is assumed.

The contemporary technique of commutation permits the establishment for each division of information of not one, but several parallel "addresses". If, for example, it is known that given information at the same time is of interest to chemistry, to physics, to biology, as well as to mensuration technique, then it is possible that this information will be recorded in an "address" under all the divisions of these fields of science connected by mutual interests.

The technical side of the question does not affect us. The difficulties here are great, but the contemporary technique is capable of coping with them.

The development of such a mechanical technique offers serious scientific and engineering interests, because the results obtained may also find wide application in automation, telemechanics and communications. The problem presented is very perspective. Because of its significance and its general character it must be solved in the Academy of Science of USSR.